SoccerTrack: A Dataset and Tracking Algorithm for Soccer with Fish-eye and Drone Videos

Atom Scott*
Ikuma Uchida*
University of Tsukuba, Japan
atom.james.scott@gmail.com
uchida.ikuma@image.iit.tsukuba.ac.jp

Masaki Onishi
National Institute of Advanced
Industrial Science and Technology, Japan
onishi@ni.aist.go.jp

Yoshinari Kameda
Kazuhiro Fukui
University of Tsukuba, Japan
{kameda@ccs, kfukui@cs}.tsukuba.ac.jp

Keisuke Fujii
Nagoya University, Japan
RIKEN/JST PRESTO, Japan
fujii@i.nagoya-u.ac.jp

Figure 1. Overview of our dataset and tracking results. We provide a dataset of (a) calibrated 8K fish-eye (wide-view) camera, (b) 4K bird-view drone camera, and (c) global navigation satellite system (GNSS) data .

Abstract

Tracking devices that can track both players and balls are critical to the performance of sports teams. Recently, significant effort has been focused on building larger broadcast sports video datasets. However, broadcast videos do not show the entire pitch and only provides partial information about the game. On the other hand, other camera perspectives can capture the whole field in a single frame, such as fish-eye and bird-eye view (drone) cameras. Unfortunately, there has not been a dataset where such data has been publicly shared until now. This paper proposes SoccerTrack, a dataset set consisting of GNSS and bounding box tracking data annotated on video captured with a 8K-resolution fish-eye camera and a 4K-resolution drone camera. In addition to a benchmark tracking algorithm, we include code for camera calibration and other

* Both authors contributed equally to this research.

preprocessing. Finally, we evaluate the tracking accuracy among a GNSS, fish-eye camera and drone camera data. SoccerTrack is expected to provide a more robust foundation for designing MOT algorithms that are less reliant on visual cues and more reliant on motion analysis. The dataset and related project code is available at https://github.com/AtomScott/SoccerTrack [1][2].

1. Introduction

In sports, fine-grained tracking data is utilized to build advanced metrics and provide teams an analytical edge. Therefore tracking systems are becoming essential for strengthening sports clubs. Recently, significant efforts have been made to build larger broadcast sports video datasets (e.g., [9, 14, 15]) and algorithms that utilize a large amount of data. For example, action spotting [5, 14, 15] and video summarization datasets [36, 37] have been released and studied for commercial usage in large markets such as the Big Five European Soccer Leagues (England Premier League, La Liga, Ligue 1, Bundesliga and Serie A) [41]. However, broadcast videos do not show the entire pitch and only provides partial information about the game. Also, it is mostly only available to wealthy professional sports teams. Therefore, teams with fewer resources find it difficult to benefit from.

Modern measurement sensors, such as the global navigation satellite system (GNSS) and local positioning systems (LPS), have enabled us to obtain players' location data on a soccer field. However, these sensors have different advantages and disadvantages (see Section 2) and are not always available due to the environment or budgets. Pioneering work using such data provided player location data by using video cameras [11] or using LPS, and 2 K panorama videos [32]. GNSS, including global positioning system (GPS), has also been intensively used mainly for conditioning [16] athletes. In other work, various indicators (e.g., moving distance) were compared with those of LPS [2] and camera-based tracking systems [33].

Camera perspectives that capture the whole field in a single frame, such as fish-eye and bird-eye view (drone) cameras, are viable options for estimating player and ball tracking data. Although high-resolution fish-eye cameras and drones may be expensive now, with the rapid growth of technology, we can expect lower-cost alternatives in the future. Several researchers [12, 18, 22] have utilized drone footage to aid in tracking from a bird's eye view, but the drone camera data has not been made available. Unfortunately, there has never been a publicly accessible dataset of annotated fish-eye and drone videos. Additionaly, the necessary preprocessing steps such as camera calibration and field registration and tracking algorithms required for camera-based tracking (see Fig. 5) using such video data is also rarely made public.

In this paper, we introduce *SoccerTrack*, a dataset set consisting of video captured with an 8K-resolution fish-eye camera and a 4K-resolution drone camera. Each video has bounding boxes annotated and is equipped with GNSS tracking data. We summarize the novelty of this paper in Table 1. Code for the core algorithms will be released, which include calibration, field registration, and multi-object tracking (MOT), as illustrated in Fig. 2. Furthermore, we will publicly share their benchmark algorithms.

The contributions of this paper are as follows.

- We build a new soccer tracking dataset called SoccerTrack, including data from fish-eye and drone cameras annotated with bounding boxes and pitch coordinates as described in Table 1.
- We propose and will share algorithms for camera calibration, tracking (players and ball) and other preprocessing as illustrated in Fig. 2.
- We perform comprehensive evaluations of the tracking accuracy between the GNSS, fish-eye and drone cameras data.

The remainder of this paper is organized as follows. First, in Section 2, We provide an overview of the related work. Next, we describe the SoccerTrack dataset in Section 3. In Section 4, we describe the tracking algorithms including calibration and other preprocessing. Then, we present the experimental results in Section 5, and conclude this paper in Section 6.

2. Related work

2.1. Multi-object tracking dataset

MOT is a well-established task in computer vision. The main objective of MOT is to track the trajectories of a collection of objects, such as pedestrians, while recognizing their identities as they move through a sequence of video frames. Many MOT datasets focused on various scenarios have been proposed. MOTChallenge [10, 25, 29] is the most widely used benchmark for monitoring multiple objects. It includes, among other things, some of the largest datasets for pedestrian tracking that are currently available to the public.

In the sports domain, pioneering work in soccer tracking datasets provide player location data with video cameras [11], and 2 K panorama videos with LPS data [32]. For other purposes, large broadcast videos (e.g., [9, 14, 20]) and event datasets (e.g., [31]) have been shared. Virtual environments such as Google research football (GFootball) [23]

[1]Project Website: https://atomscott.github.io/SoccerTrack
[2]Code Documentation: https://soccertrack.readthedocs.io/en/latest/

Dataset	Camera	Wide-view	Top-view	GNSS/LPS	Location data	Bounding box	Tracking code
D'Orazio et al. [11]	✓	✗	✗	✗	✓	✗	✗
Pettersen et al. [32]	✓	Panorama	✗	LPS	✓	✗	✗
Pappalardo et al. [31]	✗	✗	✗	✗	✓	✗	✗
GFootball [23]	✓	—	—	—	✓	✗	✗
SoccerNet v1 [14]	✓	✗	✗	✗	✗	✗	✗
SoccerNet v2 [9]	✓	✗	✗	✗	✓	✓	✓
SoccerTrack (ours)	✓	Fish-eye	Drone	GNSS	✓	✓	✓

Table 1. Overview of various representative soccer datasets. Because GFootball [23] is a virtual environment, the camera types cannot be defined. In SoccerNet v2 [9], location data and tracking codes were available at the Tracking Challenge (a competition).

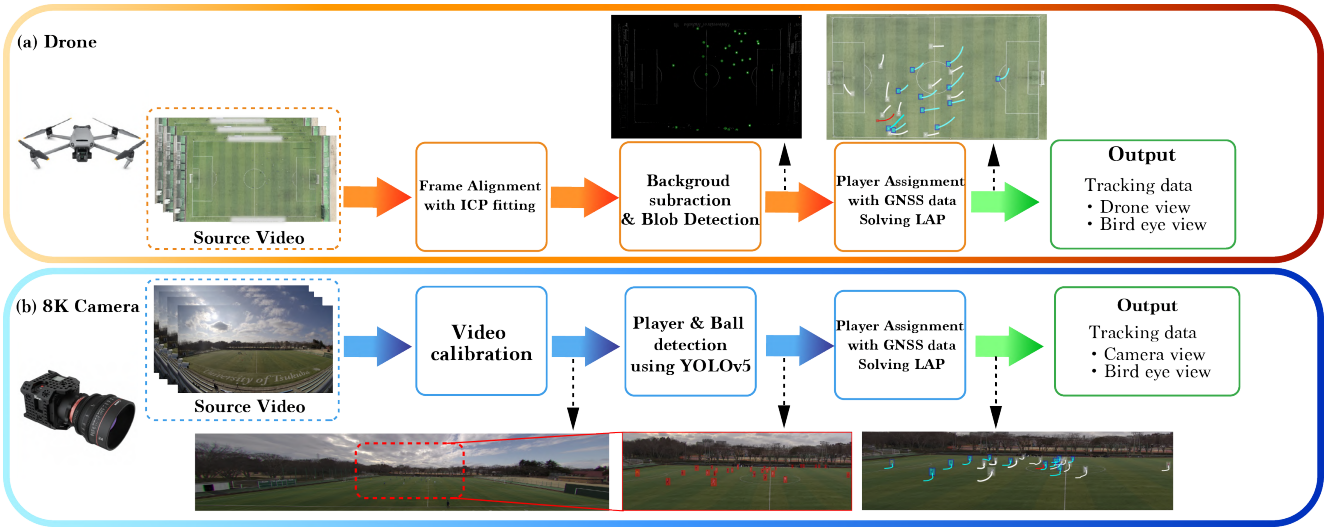

Figure 2. The process for building a dataset. (a) Using a 4K drone video data, we perform frame alignment with iterative closest point (ICP) fitting, background subtraction and blob detection. (b) Using a 8K fish-eye video data, we perform video calibration, and player and ball detection using YOLOv5 [21]. Thereafter, for both data, we perform player alignment with GNSS data solving a linear assignment problem (LAP) and obtain tracking data from bird-eye and wide-view.

can also be used to generate synthetic camera and location data. In other team sports, basketball [6, 27] and volleyball [19] video datasets have been shared publicly.

These measurement systems have different pros and cons. GNSS and LPS have advantages in that they do not require tracking algorithms, but they require sensors to be attached and have basically worse spatio-temporal resolution than camera systems. The recent development of GNSS, which comprises multiple tracking satellite systems such as the GPS, GLONASS, Galileo and BeiDou, has improved the availability and signal strength of surrounding satellites compared with traditional GPS devices [8]. A previous study investigated the GNSS accuracy of velocity and acceleration using motion capture data [8] and that of the running distance [4]; however, the position accuracy was unknown and the data were not shared. GNSS/GPS is easier to use but has a worse spatio-temporal resolution than LPS [17, 35]. Fish-eye and panorama cameras can capture the full pitch using one and multiple cameras, respectively. The fish-eye video saves human labor costs; however, the image is sometimes distorted and requires calibration. Drone cameras can provide more accurate 2D coordinates without occlusion if they can capture the image from bird-view; however, they require another source of information to identify the players. In this paper, we provide all three sources (fish-eye, drone and GNSS) of data and annotated bounding boxes for fish-eye and drone videos.

2.2. Multi-object tracking algorithms

The typical approach to MOT algorithms follows the tracking-by-detection paradigm, which attempts to solve the problem in two steps. (1) The detection model detects items of interest via bounding boxes in each frame, then (2) the association model extracts visual re-identification (re-ID) features corresponding to each bounding box, ties the detection to an existing track, or generates a new track based on specified metrics set on features. Such approaches benefit from increasingly powerful image recognition backbones

to extract important visual features.

Scalability is a challenge for tracking-by-detections MOT algorithms. When there are a high number of objects in the environment, the inference speed decreases because the two models do not integrate features and must apply the re-ID models for each bounding box individually in the video. Recent advances in multi-object tracking have centered on the joint-detection-and-tracking paradigm, in which object localization and association are performed concurrently [28, 30, 43, 49, 52, 53].

Most appearance-based tracking paradigms have a fatal flaw in instances when objects have highly similar appearances, for as in team sports where members of the same team wear identical gear. For example, DeepSORT [50] attempts to combine distance measures based on motion states with deep appearance descriptors, however when tuned for performance, the resulting parameters depend significantly on appearance over motion. The authors of DanceTrack [42] advocate for a more comprehensive and intelligent tracking system that incorporates additional inputs into modeling, such as object motion patterns and temporal dynamics. DanceTrack is a large-scale dataset that emphasizes tracking targets with uniform appearance and diverse motion. Although it is a significant effort to counter the recent appearance-focused paradigm, an algorithm that highly outperforms previous methods using motion analysis was not presented. The proposed dataset also lacks depth information, which is crucial for fine-grained motion.

We provide the first transparent baseline for player position estimation with interchangeable modules, that relies on modern techniques and freely available data, while evaluating each module.

2.3. Application of tracking data

The positional coordinates of players on the soccer pitch are fundamental information in soccer analysis and tactical understanding. Using Tracking data, one can analyze various aspects of soccer. For example, event detection, such as automatic offside detection [46], and movement evaluations, such as shots and passes, (e.g., [7, 38]), off-ball (e.g., [40]) and defense (e.g., [45]) can be realized. Tracking data will enable the simulation of future motions via trajectory prediction [13, 24, 44]. It also enables applications such as trajectory similarity and retrieval [47, 48]. GNSS/GPS has been intensively used mainly for conditioning [16] and compared with LPS [2] and camera-based tracking systems [33]. However, only performance indices such as distance and velocity were compared in these studies, and the data and tracking accuracy for both sources have not been evaluated and publicly shared.

Figure 3. GNSS devices used in the measurement experiment. Athletes inserted the device inside a special vest to prevent the device from falling.

3. SoccerTrack dataset

In this section, we describe the MOT tracking dataset, *SoccerTrack*. We explain the dataset construction process, dataset structure and evaluation metrics, in that order.

3.1. Dataset Construction

All data in SoccerTrack was obtained from 11-vs-11 soccer games between college-aged athletes. Measurements were conducted after we received the approval of Tsukuba university's ethics committee, and all participants provided signed informed permission. After recording several soccer matches, we annotated the videos semi-automatically based on the GNSS coordinates of each player.

We illustrate the full dataset construction procedure in the following four steps:

(i) Capture fish-eye and bird-eye view (drone) video and perform camera calibration;

(ii) Transform GNSS data collected from wearable devices into the pitch coordinates;

(iii) Synchronize the time between the drone video and GNSS data, and assign IDs to annotated bounding boxes.

(iv) Synchronize the time between the drone video and fish-eye video, and assign IDs to annotated bounding boxes.

An overview of the SoccerTrack dataset is provided in Table 2. Next, we describe the details of each step listed above.

3.1.1 Video Collection

We used a fish-eye lense (Z CAM E2-F8, China) to enlarge the camera's field of view and adjusted the position of the camera beforehand to capture the entire soccer field. The

	Wide-view camera	Top-view camera	GNSS
Device	Z CAM E2-F8	DJI Mavic 3	STATSPORTS APEX 10 Hz
Resolution	8 K (7,680 × 4,320 pixels)	4 K (3,840 × 2,160 pixels)	Abs. err. in 20-m run: 0.22 ± 0.20 m [4]
FPS	30	30	10
Player tracking	✓	✓	✓
Ball tracking	✓	✓	✗
Bounding box	✓	✓	—
Location data	✓	✓	✓
Player ID	✓	✓	✓

Table 2. Overview of SoccerTrack dataset.

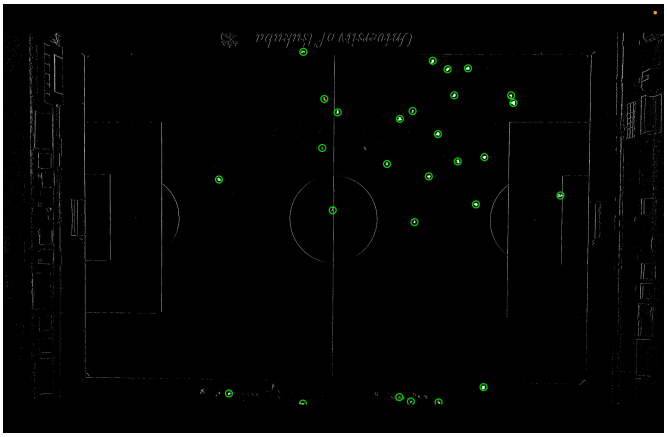

Figure 4. An example of detected blobs. Using a foreground mask to detect moving objects in the frame enables the detection of objects from the drone viewpoint video without learning.

resolution was 8 K (7,680 × 4,320 pixels), the frame rate was 30 fps, and the raw video data was approximately 1 TB. To remove lens distortion from the fish-eye video, we recorded a checkerboard calibration pattern from various angles to calculate the camera's intrinsic and extrinsic parameters. We implemented a k-means clustering algorithm to select a diverse set of checkerboard images automatically. As a result, 200-300 frames were selected and used as input to OpenCVs fish-eye calibration method.

To record from a bird's-eye perspective, we used a DJI Mavic drone (Da-Jiang Innovations Science and Technology Co., Ltd., China). The resolution was 4 K (3,840 × 2,160 pixels), frame rate was 30 fps, and raw video data was approximately 250 GB.

3.1.2 GNSS Data Collection

GNSS (STATSports Apex 10 Hz, Northern Ireland) data were collected outdoors on the soccer field. The STATSports Apex unit is an athlete-tracking system released in August 2017, which is widely used in professional clubs (e.g., in the Premier League and Serie A). During measurement, the players wore the device in a custom-made sports vest. The GNSS simultaneously acquires and tracks multiple satellite systems (e.g., GPS, GLONASS, Galileo, and BeiDou), thus providing more accurate positional information than typical GPS devices.

GNSS data preprocessing was performed as follows. First, we obtained each player's latitude and longitude coordinates from the GNSS devices. Next, this data was then projected onto pitch coordinates via a homography transform derived by manually annotating keypoints in the geographic coordinate system. As a result, pitch coordinates are normalized to be within a range between 0-68 × 0-105 m. Finally, after visually inspecting the GNSS coordinates to confirm no apparent outliers, we project the GNSS coordinates onto the drone camera space to prepare for semi-automatic annotation.

A previous study [4] evaluated the accuracy of the same GNSS devices. They used three courses, a 400-m athletic track, a custom team-sports oriented circuit of 128.5-m, and a 20-m sprint, for evaluation. Results show that the accumulated absolute errors from the running distance were 4.19 ± 3.48 m, 2.85 ± 1.4 m, and 0.22 ± 0.20 m, respectively. For evaluation of GNSS accuracy, we followed [2] using two selected course tests; (i) linear course and (ii) circular course. The trials were performed at two speeds: walking (< 6 km/h) and running (> 16 km/h). A bias-corrected root mean square error (RMSE) score was calculated using the euclidean distance between the course and the GNSS measurement. The GNSS coordinates were are projected to a pitch plane (as explained in Section 3.1.3) so that RMSE can be interpreted in meters. We performed a grid search to choose optimal offsets to use as bias. The results are shown in Table 3. We found similar tendency to the result of a GPS study [2], which showed the accuracy was slightly more affected by the movement speed.

Course	Speed	Bias	Ave. RMSE* (std.)
Linear	Walk	-1.970 -1.162	0.514 (0.183)
Linear	Run	-1.263 -1.060	1.163 (0.157)
Circular	Walk	-1.465 -1.667	0.436 (0.456)
Circular	Run	-1.060 -1.354	0.769 (0.140)

Table 3. Bias and average RMSE of performed trials. Two bias values are about x and y coordinates. Average RMSE is calculated after bias correction.

3.1.3 Drone Video Annotation

This section details our approach to efficiently annotating bounding boxes in the drone video. Manual annotation is a laborious task that takes around between five and six hours to annotate one minute of footage. We quickly concluded that annotating all footage would be impracticable. Instead, we utilized well-established computer vision techniques to determine reasonable estimates to shorten the process. Fig. 2 (a) is a pipeline showing the semi-automatic annotation process on a drone.

First, we used maximum contour detection after line extraction on binarised frames. We then adjusted the line length settings to ensure that the maximum contour always resembled the outer lines of the soccer field. To fill holes in the pitch lines, preprocessing operations such as Gaussian blurring and morphological closing. The Douglas–Peucker algorithm [34] was applied to simplify the resulting maximum contour into a rectangle which was then used as Region of Interest (ROI). Next, we detect blobs in foreground masks generated with the K-Nearest Neighbor (KNN) background subtraction algorithm as shown in Fig. 2 [56]. Blobs are found using the Determinant of the Hessian (DoH) method [3]. Both algorithms were chosen above others after a reasonable balance between speed and accuracy was demonstrated. After removing blobs outside the ROI, the remaining blobs were projected onto the pitch plane in a manner similar to that described in Section 3.1.2. These calibration and preprocessing are used in our tracking system.

Once the blobs and GNSS coordinates are on an identical coordinate system, we formulate a Linear Assignment Problem to assign a GNSS device ID to each blob. Prior to assignment, iterative closest point (ICP) [1] was employed to minimise any discrepancy introduced by preprocessing, GNSS sensor error, and other sources of inaccuracy.Finally, we import the assignment data into the Computer Vision Annotation Tool (CVAT) [39] to manually inspect and correct false annotations. CVAT provides a simple mechanism that linearly interpolates bounding boxes between missing annotations. By combining linear interpolation and estimations based on GNSS-ID assign blobs, we were able to significantly speedup the annotation process. After annotating 30 minutes worth of data with this method, we replaced the DoH blob detector with a fine-tuned YOLOv5 [21] object detector pre-trained on the COCO object detection dataset [26] as it was more accurate at detecting the players. We used a similar procedure to train an object detector for ball detection, except that for the first 30 minutes, we manually annotated all frames without using any image processing algorithms.

3.1.4 Fish-eye Video Annotation

To perform annotation on fish-eye videos, we manually annotated keypoints of a single frame and computed a homography matrix to transform coordinates from fish-eye frames to pitch coordinates. Since the fish-eye camera is fixed, we did only need a single homography matrix. Next, we perform object detection using a pre-trained YOLOv5 object detector. Unlike the drone videos, YOLOv5 was able to detect the most of the players without fine-tuning. We then estimated the ID of each bounding box in each frame by performing ICP between detections in fish-eye videos and annotated drone videos. Since the YOLOv5 object detector was not able to detect every player perfectly, a number of estimated IDS were incorrect. These incorrect assignments were fixed manually in CVAT.

3.2. Dataset Structure

Here we outline the structural information of the SoccerTrack dataset. At the time of writing, this dataset consists of 20 clips of 30-second video footage (.mp4) captured by fisheye and drone cameras. Each clip is accompanied by an annotation file in simple comma-separated value (CSV) format. Each line of the csv file represents one object instance and contains the values; $frame$, id, bb_left, bb_top, bb_width, bb_height. We also provide corresponding GNSS data in CSV format. Manually annotated pitch key-point coordinates are stored in JSON files. The first 15 video clips (7.5 minutes) will be utilized for training and tuning model parameters, while the last 5 clips (2.5 minutes) will be used as a testset.

3.3. Evaluation Metrics

Multi-Object Tracking (MOT) has a reputation for being tough to evaluate correctly. While metrics such as the commonly used MOTA tend to overemphasize accurate detection, IDF1 and AssA will, on the other hand, overemphasize association quality. Therefore, Higher-Order Tracking Accuracy (HOTA) has been adopted as the primary metric in

several recent benchmarks since its proposal (BDD100K, KITTI, DanceTrack). HOTA aims to balance detection and association by explicitly combining a DetA and AssA. However, we were not able to find simple implementations of the HOTA metric. In contrast to the standard MOT setting, where the number of tracking targets is unknown, we know that there is one ball and 22 players in soccer. Although this characteristic may open opportunities to custom metrics particular to n-known object tracking problems, we chose to abide by current MOT practices. Thus, we use MOTA as the key performance indicator for SoccerTrack. In addition, we also provide statistics regarding False Positives (FP), and ID switches (IDs).

3.4. Limitation

Here we discuss some of the known limitations of the SoccerTrack dataset. Although we significantly improve over past broadcast view datasets by providing a fish-eye and birds-eye view in addition to bounding box labels and GNSS coordinates, we acknowledge the following shortcomings. To begin, we were unable to provide a diversity of environmental conditions. For instance, all games were played on a single field, and there are only two distinct sets of soccer jerseys. Furthermore, since the weather was sunny primarily, trackers trained on this dataset may not perform well in the rain. Additionally, we could not conduct experiments with fine-grained data such as human pose or segmentation masks, which are sure to be of interest to sports science experts. Most of all, we could only provide a fraction annotated data in this release due to time and resource constraints. Out of approximately 120 minutes of recorded video data, we annotated 10 minutes of both drone and fish-eye video. We continue to work on the remaining data and plan to release a fully annotated dataset by the end of the current year. To look on the bright side, our efforts were received with warm support and we are given future chances to measure more games at several venues. We are optimistic about providing more comprehensive datasets in the future.

4. SoccerTrack Algorithm

We provide tracking algorithms including camera calibration and other preprocessing procedures for the SoccerTrack dataset. We use camera calibration and other preprocessing procedures for drone and fish-eye camera data as described in Sections 3.1.3 and 3.1.4, respectively. As our tracking algorithm, we extend [46]'s tracker, which is essentially a modified version of DeepSORT. DeepSORT uses a combination of motion and appearance metrics to assign detections to tracklets. The following subsections outline the procedures involved in metric calculation and assignment.

4.1. Motion Information

Our motion model follows the modifications in [46], where the authors use coordinates projected onto the pitch plane as the observation input of the motion model. In contrast, DeepSORT directly uses the bounding box coordinates from the object detector. In addition to being more intuitive, pitch coordinates are the de facto representation for many sophisticated motion models [24, 51]. Therefore, although we adopt a basic Kalman Filter model, we anticipate future work to improve on it in order to improve tracking accuracy and efficiency. Further, we introduce simple spatial constraints to eliminate tracklets that are out-of-bounds.

4.2. Visual Information

We use the omni-scale network (OSNet) [55] architecture to extract deep appearance features from detected bounding boxes. An OSNet is pre-trained on a large-scale person re-identification dataset [54] for use in fish-eye view camera. On the other hand, we train an OSNet from scratch for use in drone video. This is due to the fact that top-view images are significantly different from images in the large-scale person re-identification dataset. Our training approach is not covered in depth in this work, but we will make a pre-trained model available in our GitHub repository.

4.3. Assignment

An effective method of solving the association problem between existing tracks and newly acquired detections in the detect-then-track paradigm is to construct an assignment problem that can be solved using the Hungarian algorithm. DeepSORT combines a motion descriptor and a deep appearance descriptor into a single cost function with the goal of minimizing the total assignment cost. In similar fashion, we combine the two metrics explained in Section 4.1 and Section 4.2 The influence of each metric on the combined association cost can be controlled through hyperparameter λ. We select the best λ by performing grid search over the training set.

5. Experiments

In this section, we perform two experiments: evaluations of location data accuracy and tracking algorithm performance. First, we evaluate location accuracy of the Fish-eye video and GNSS data compared with the drone video, which has less distortion or bias because of the top-view video. Second, we assess the performance of our tracker on both fish-eye and drone videos. The primary evaluation metric used is MOTA, also we mention a few other complimentary metrics, as explained in Section 3.3. In the most common use cases of MOTA, the link between ground truth objects and tracker output is established by intersection over union (IoU) with a threshold of 0.5 as similarity

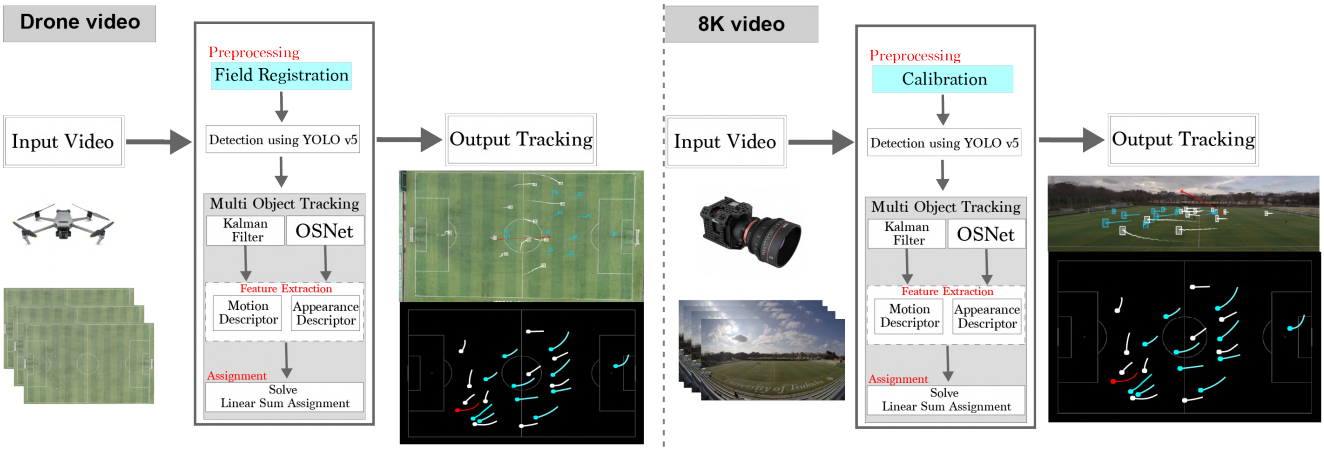

Figure 5. The pipeline of the tracking system. By building an algorithm inspired by DeepSORT, both visual and motion information are taken into account. Field Registration is performed for drone video (left), and calibration is performed for 8K fish-eye video as preprocessing of video (right).

criteria. However, since the drone bounding boxes present in the drone video are significantly smaller than those in the usual MOT setting, we lower this threshold to 0.2. The last 1000 frames were used to test each camera view, while the rest of the frames were used to train and select hyperparameters.

5.1. Location data accuracy

First, we evaluated the keypoint accuracies in the drone, fish-eye, and GNSS. We used 65 key points (for details, see shared code) and computed L2 errors (mean $\pm$ standard deviation [m]) between the ideal and estimated key points in the soccer field for the three sources. Results show that the errors in the drone camera (0.06 ± 0.03) and GNSS projection error (0.13 ± 0.09) were smaller than that in the fish-eye (0.56 ± 0.42). Next, similarly to Section 3.1.2, a bias-corrected RMSE was calculated using the euclidean distance between the GNSS and drone data and that between the fish-eye and drone data. We assume that the drone data has less distortion or bias because of the top-view video and use them as the ground truth. First, a homography matrix estimated from the manually annotated keypoints projected all bounding box annotations to pitch coordinates. We then upsampled the GNSS data (10 Hz) to 30 Hz (drone and fish-eye data) and computed temporal mean values of the bias-corrected RMSE (3320 frames) for each player (22 players in total). Results show that the bias-corrected RMSE (mean $\pm$ standard deviation [m] among 22 players) of location data in the fish-eye camera (2.76 ± 2.86) was similar to (but the SD was smaller than) that in the GNSS (2.77 ± 4.47) but that of the velocity data in the fish-eye camera (2.14 ± 0.51) was more accurate than that in the GNSS (2.62 ± 0.34). Overall, the accuracy of the fish-eye camera annotations was better than that of GNSS. Although the GNSS accuarcy was already investigated [17, 35], little research has been done on the fish-eye camera accuracy. The fish-eye camera can capture finer interactive movements of players and a ball.

5.2. Tracking performance

We performed tracking using both fish-eye and drone test videos (five 30-second clips for each camera view). All metrics were calculated for each clip and then averaged. The average tracking performance in fish-eye video resulted in a 14.2% MOTA score, 19691 FPs, and 19 ID switches. The average tracking performance in the drone video resulted in a 57.50% MOTA score, 8781 FPs, and 5 ID switches. We observe that the low MOTA score for fisheye camera may be because of falsely tracked people (such as coaches or bench-starters). The large number of false negatives is mostly due to low confidence object detection results and can be decreased by tuning confidence thresholds or using a better object detector.

6. Conclusion

In this paper, we presented SoccerTrack, the first fish-eye and birds-eye view video dataset for detection and tracking in soccer. We experimentally demonstrated that both views could be used to perform tracking. We anticipate that our research will pave the way for future monitoring initiatives in a variety of other sports, in addition to soccer.

7. Acknowledgement

This work was supported by JSPS KAKENHI (Grant Number 20H04075) and JST Presto (Grant Number JPMJPR20CA).

References

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987. 6

[2] Alejandro Bastida Castillo, Carlos D Gómez Carmona, Ernesto De la Cruz Sánchez, and José Pino Ortega. Accuracy, intra-and inter-unit reliability, and comparison between gps and uwb-based position-tracking systems used for time–motion analyses in soccer. *European journal of sport science*, 18(4):450–457, 2018. 2, 4, 5

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 6

[4] Marco Beato, Giuseppe Coratella, Adam Stiff, and Antonio Dello Iacono. The validity and between-unit variability of gnss units (statsports apex 10 and 18 hz) for measuring distance and peak speed in team sports. *Frontiers in physiology*, 9:1288, 2018. 3, 5

[5] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2020. 2

[6] Christophe De Vleeschouwer, Fan Chen, Damien Delannay, Christophe Parisot, Christophe Chaudy, Eric Martrou, Andrea Cavallaro, et al. Distributed video acquisition and annotation for sport-event summarization. *New European Media Summit*, 8, 2008. 3

[7] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1851–1861, 2019. 4

[8] Jace A Delaney, Taylor M Wileman, Nicholas J Perry, Heidi R Thornton, Mark P Moresi, and Grant M Duthie. The validity of a global navigation satellite system for quantifying small-area team-sport movements. *The Journal of Strength & Conditioning Research*, 33(6):1463–1466, 2019. 3

[9] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4508–4519, 2021. 2, 3

[10] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. CVPR19 tracking and detection challenge: How crowded can it get? *arXiv:1906.04567 [cs]*, June 2019. arXiv: 1906.04567. 2

[11] Tiziana D'Orazio, Marco Leo, Nicola Mosca, Paolo Spagnolo, and Pier Luigi Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564. IEEE, 2009. 2, 3

[12] Filipe Trocado Ferreira. Video analysis in indoor soccer with a quadcopter. Master's thesis, Mestrado Integrado em Engenharia Eletrotécnica e de Computadores, 2014. 2

[13] Keisuke Fujii, Naoya Takeishi, Yoshinobu Kawahara, and Kazuya Takeda. Policy learning with partial observation and mechanical constraints for multi-person modeling. *arXiv preprint arXiv:2007.03155*, 2020. 4

[14] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018. 2, 3

[15] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2021. 2

[16] Liam Hennessy and Ian Jeffreys. The current use of gps, its potential, and limitations in soccer. *Strength & Conditioning Journal*, 40(3):83–94, 2018. 2, 4

[17] Matthias W Hoppe, Christian Baumgart, Ted Polglaze, and Jürgen Freiwald. Validity and reliability of gps and lps for measuring distances covered and sprint mechanical properties in team sports. *PloS one*, 13(2):e0192708, 2018. 3, 8

[18] Viatcheslav Iastrebov, Choon Yue Wong, Wee Ching Pang, Gerald Seet, V Iastrebov, CY Wong, WC Pang, and G Seet. Motion tracking drone for extreme sports filming. In *1st international conference in sports science & technology (ICSST 2014). https://hdl. handle. net/10356/138183*, 2014. 2

[19] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[20] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. Soccerdb: A large-scale database for comprehensive video understanding. In *Proceedings of the 3rd International Workshop on Multimedia Contentf Analysis in Sports*, pages 1–8, 2020. 2

[21] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020. 3, 6

[22] Stephen Karungaru, Kenji Matsuura, Hiroki Tanioka, Tomohito Wada, and Naka Gotoda. Ground sports strategy formulation and assistance technology develpoment: player data acquisition from drone videos. In *2019 8th International Conference on Industrial Technology and Management (ICITM)*, pages 322–325. IEEE, 2019. 2

[23] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environ-

ment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4501–4510, 2020. 2, 3

[24] Hoang M Le, Peter Carr, Yisong Yue, and Patrick Lucey. Data-driven ghosting using deep imitation learning. In *Proceedings of the 11th MIT sloan sports analytics conference*, 2017. 4, 7

[25] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[27] Keyu Lu, Jianhui Chen, James J Little, and Hangen He. Light cascaded convolutional neural networks for accurate player detection. In *British Machine Vision Conference (BMVC)*, 2017. 3

[28] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. 2021. 4

[29] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 2

[30] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021. 4

[31] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):1–15, 2019. 2, 3

[32] Svein Arne Pettersen, Dag Johansen, Håvard Johansen, Vegard Berg-Johansen, Vamsidhar Reddy Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Kvale Stensland, and Pål Halvorsen. Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 18–23, 2014. 2, 3

[33] Eduard Pons, Tomás García-Calvo, Ricardo Resta, Hugo Blanco, Roberto López del Campo, Jesús Díaz García, and Juan José Pulido. A comparison of a gps device and a multi-camera video technology during official soccer matches: Agreement between systems. *PloS one*, 14(8):e0220729, 2019. 2, 4

[34] Urs Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing*, 1(3):244–256, 1972. 6

[35] Markel Rico-González, Asier Los Arcos, Filipe M. Clemente, Daniel Rojas-Valverde, and José Pino-Ortega. Accuracy and reliability of local positioning systems for measuring sport movement patterns in stadium-scale: A systematic review. *Applied Sciences*, 10(17), 2020. 3, 8

[36] Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. Hierarchical multimodal attention for deep video summarization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7977–7984. IEEE, 2021. 2

[37] Melissa Fallas Sanabria, Frédéric Precioso, and Thomas Menguy. Profiling actions for sport video summarization: An attention signal analysis. *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2020. 2

[38] Atom Scott., Keisuke Fujii., and Masaki Onishi. How does ai play football? an analysis of rl and real-world football strategies. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,*, pages 42–52. INSTICC, SciTePress, 2022. 4

[39] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020. 6

[40] William Spearman. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference*, pages 1–17, 2018. 4

[41] Statista. Revenue of the biggest (big five) european soccer leagues from 1996/97 to 2021/22. (Accessed: 2022/2/21). 2

[42] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. *arXiv preprint arXiv:2111.14690*, 2021. 4

[43] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv: 2012.15460*, 2020. 4

[44] Masakiyo Teranishi, Keisuke Fujii, and Kazuya Takeda. Trajectory prediction with imitation learning reflecting defensive evaluation in team sports. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 124–125. IEEE, 2020. 4

[45] Kosuke Toda, Masakiyo Teranishi, Keisuke Kushiro, and Keisuke Fujii. Evaluation of soccer team defense based on prediction models of ball recovery and being attacked: A pilot study. *PloS One*, 17(1):e0263051, 2022. 4

[46] Ikuma Uchida, Atom Scott, Hidehiko Shishido, and Yoshinari Kameda. Automated offside detection by spatio-temporal analysis of football videos. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 17–24, 2021. 4, 7

[47] Zheng Wang, Cheng Long, and Gao Cong. Similar sports play retrieval with deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 4

[48] Zheng Wang, Cheng Long, Gao Cong, and Ce Ju. Effective and efficient sports play retrieval with deep representation learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 499–509, 2019. 4

[49] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *The European Conference on Computer Vision (ECCV)*, 2020. 4

[50] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 4

[51] Raymond A Yeh, Alexander G Schwing, Jonathan Huang, and Kevin Murphy. Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4610–4619, 2019. 7

[52] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 4

[53] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 4

[54] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer, 2016. 7

[55] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019. 7

[56] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006. 6